Streaming Data Cleaning and Probing

Abstract Raw data, especially streaming data, is usually dirty in terms of data syntactics, data semantics and data coverage. In the process of data cleaning, users get to know the basics of the dataset, such as the completeness of the data, the scales, the attributes and the value distributions. We define such effort as **data probing**. There are three phases in data cleaning, data diagnostics, data transformation and cleaning workflow reusability/verification. Existing interactive data cleaning tools concern mainly data transformation and reusability. But I argue that without **data diagnostics** users cannot fully detect all the flaws of the underlying dataset, neither can they achieve a more comprehensive understanding of the dataset. In this document, I proposed **a data cleaning and probing framework**, which visually chaines transformation units (filter units) in series into workflow. The filter unit integrates data diagnotics, data transformation and data verification together for each transformation step. Users can adapt the original or modified workflow to incoming data or other dataset. Based on the framework, I want to implement a B/S system that runs locally, and supports large scale streaming data cleaning and probing.
Key words streaming data; data cleaning; data probing.

Introduction

Challenges in data cleaning and probing [2-8]
data syntactics: inconsistent data schema, wrong values, invalid data format, abbreviations...
data semantics: values that are semantically wrong (e.g. criminals under 7 years old), change of data colleciton semantics (e.g. most popular women singers vs. most popular singers)...
data coverage: missing data, missing data fields, duplicate data items...

Special concern with streaming data where time is an embedded data field [1]
Domain violation: overlap in time intervals, start time > end time...
Heterogeneous sources: inconsistent time format (format and time zone), inconsistent time scale, inconsistent data collection methods...

Things have been done
Flexible raw data sources: CSV file, PDF file, URLs, web APIs, databases..
data diagnostics: Google Refine [12] (limited --- numeric or categorical values distribution)
user defined data transformation (extracting, deleting, de-duplicating, integrating): Google Refine [12], Data Wrangler [10]
schema mapping: Schema Mapper [11]
transformation script reusing: Google Refine [12], Data Wrangler [10]
transformation language: Google Refine[12], Data Wrangler [10]

Things not done yet…

more powerfule and flexible data diagnostics with visualizaiton: spatialtemporal value distribution, user defined dianosing visualization techniques (e.g. scatter plots, parallel coordinates)
diagnosing with verification: every transformation step can be verfied right after it is done (optional).
Interactive workflow: simple, interactive, reusable, editable workflow
Also, the workflow can be embedded into existing streaming data framework, which takes raw data in and produce clean data out.

Contributions

1. A framework that supports data probing and data cleaning in one reusable workflow;
2. A interactive workflow that chains transformation steps (filter units) in series;
3. Filter units that integrate visual data diagnotics, visual data transformation and visual data verification together for each transformation step.

Related Work

Data quality [1-8]
Existing tools [1--12]: Google refine, Data Wrangler, Schema Mapper

Framework Design

The overall architecture of the workflow is shown in Figure 1. First, data is imported to the cleaning and probing system, and connected to an undefined **filter unit**. Then users define the unit by specifying a visual diagnosing method, a transformation method and a verification method, as in Figure 2. Filter units can be joined in parallel or in series to construct the workflow. At the end, users can export either the clean data or the workflow.

For a specific filter unit, diagnosing method is defined first by users choosing one basic visualization method (scatter plots, parallel coordinates…), mapping corresponding attributes to visual encodings. Note that every diagnosing method may or may not diagnose flaws in the dataset. A transformation method is further specified after user diagnosing the problem. Visual verification is optional but encouraged to compare the effect of transformation. If applied, visual verification is usually visualized by the same method with diagnosing method.

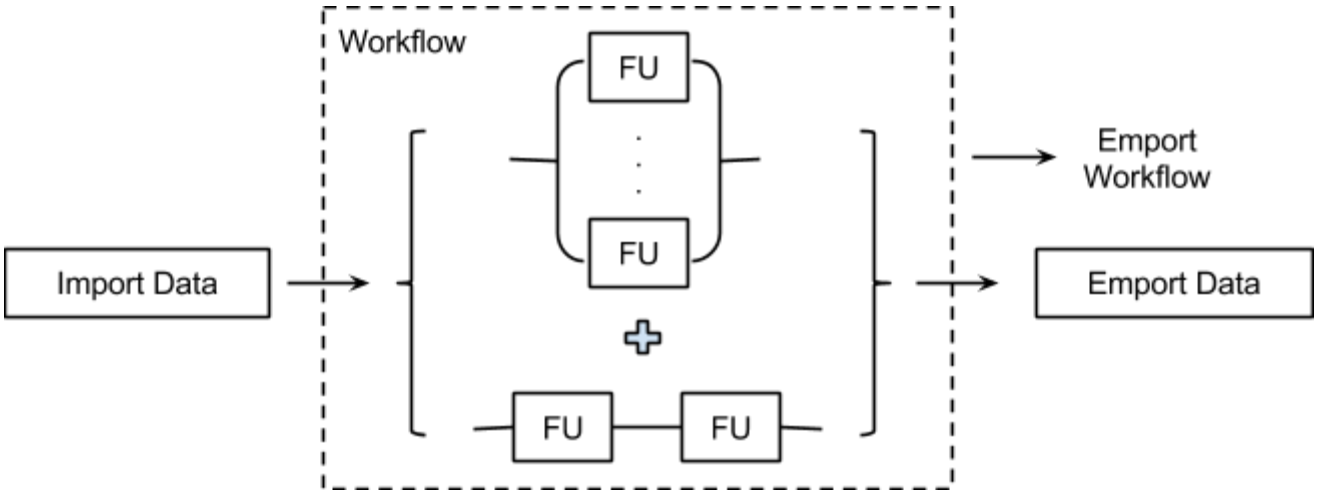

Figure 1. Framework Architecture

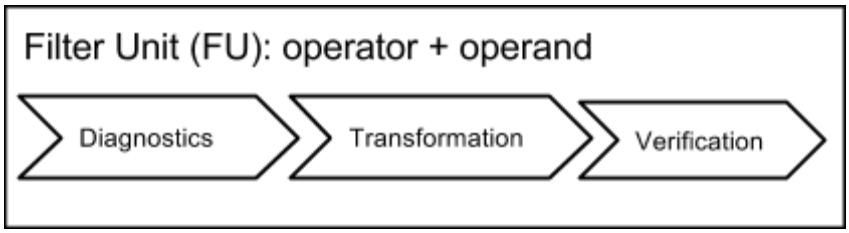

Figure 2. Filter Unit

System Implementation (still open)

data construction
database
schema language
visual display
Others

Potential Useful Datasets

1. VAST Challenge datasets --- they are large and temporal and complex, very suitable for our project
2. Fortune-500 dataset --- small dataset but very dirty, and I don't see any existing data cleaning tools can deal with this dataset because it requires a lot of domain knowledge. Good for testing joint schema function (companies to company code such as SIC or NAICS).
3. OpenStreetMaps dataset. --- Shehzad is experienced with cleaning OpenStreetMaps dataset. It's not time series but it's geospatial dataset.
4. Aliyun dataset could be a good streaming example, but we have no access to it.
5. Twitter or Valet crime? --- well structured, but they're spatial temporal.

6. Govenment datasets, such as flight delay dataset?

References

Theories of data quality and data wrangling

1) Gschwandtner, Theresia, et al. "A taxonomy of dirty time-oriented data."*Multidisciplinary Research and Practice for Information Systems*. Springer Berlin Heidelberg, 2012. 58-72.
2) Kandel, Sean, et al. "Research directions in data wrangling: Visualizations and transformations for usable and credible data." *Information Visualization*10.4 (2011): 271-288.
3) Rahm, E., Do, H.H.: Data Cleaning: Problems and Current Approaches. IEEE Techn. Bulletin on Data Engineering 31 (2000)
4) Kim, W., Choi, B.-J., Hong, E.-K., Kim, S.-K., Lee, D.: A Taxonomy of Dirty Data. Data Mining and Knowledge Discovery 7, 81–99 (2003)
5) M¨uller, H., Freytag, J.-C.: Problems, Methods, and Challenges in Comprehensive Data Cleansing. Technical report HUB-IB-164, Humboldt University Berlin (2003)
6) Oliveira, P., Rodrigues, F., Henriques, P.: A Formal Definition of Data Quality Problems. In: International Conference on Information Quality (MIT IQ Conference) (2005)
7) Barateiro, J., Galhardas, H.: A Survey of Data Quality Tools. Datenbankspektrum 14, 15–21 (2005)
8) Sadiq, S., Yeganeh, N., Indulska, M.: 20 Years of Data Quality Research: Themes, Trends and Synergies. In: 22nd Australasian Database Conference (ADC 2011), pp. 1–10. Australian Computer Society, Sydney (2011)
9)

Existing tools

10) Kandel, Sean, et al. "Wrangler: Interactive visual specification of data transformation scripts." *PART 5--------Proceedings of the 2011 annual conference on Human factors in computing systems*. ACM, 2011.
11) Robertson, George G., Mary P. Czerwinski, and John E. Churchill. "Visualization of mappings between schemas." *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 2005.
12) Huynh D and Mazzocchi S. Freebase GridWorks. http://code.google.com/p/google-refine/
13)

Potential Useful Architecture Techniques

14) Lins, Lauro, James T. Klosowski, and Carlos Scheidegger. "Nanocubes for Real-Time Exploration of Spatiotemporal Datasets." *Visualization and Computer Graphics, IEEE Transactions on* 19.12 (2013): 2456-2465. --- LoD technique, not applicable